Voice Server with Under-resourced Acoustic Model: Application to Agricultural Extension

Minh H. Le, Truong M. Tran, Nhut M. Pham, and Quan H. Vu

Artificial Intelligence Laboratory, University of Science, VNU-HCM, HCM, Vietnam

Abstract- The majority of farmers in developing countries are unable to reach farming information and knowledge, and so they often rely on rudimentary methods. To enhance the agricultural life and productivity, one must ensure the flow of information to farmers; that is to bridge an agricultural info channel to farmers effectively. In this work, we move on to build an agricultural IVR system with a mission of providing agricultural info services to the farmers. However, since the underlying techniques resort to statistical approaches, a large amount of training material is demanded to reach a sustainable level of performance, especially on a new applied domain. Unfortunately, the required effort to develop such corpora is both costly and time consuming, and large scale acquisition campaigns might not be feasible. Under these restricted circumstances, subspace Gaussian mixture models and parameter synthesis are the keys to build ASR and TTS components of an IVR system. Experimental results confirm the hypothesis with 3.43% winning over conventional methods in an application of agricultural extension.

I. Introduction

The telephone — whether landline or mobile — is often the handiest, user-friendliest access device. As a result, people can increase access to their business services and applications by existing speech-enabling applications [1]. With the support achieved from spoken language processing (SLP) techniques, these applications can be promoted into a new type of human interaction: interactive voice response (IVR) [1]. Users can interact with an IVR system (technically called "voice server") as if it were a conversational partner.

However, telephone-based interactions pose several research challenges [2]. For example, telephone speech is often hard to recognize and understand due to the reduced channel bandwidth and the presence of noise. In addition, voice-based interaction relies on only the human auditory channel to receive the information, and thus potentially increases the cognitive load. Furthermore, real-time performance is necessary, since prolonged delay over the phone can be quite annoying to users and render the system unusable.

Voice Server has been insensitively researched for a long time. In 1997, Victor Zue et al [2] had begun to develop JUPITER, a conversational interface that allows users to access and receive on-line weather forecast information for over 500 cities worldwide over the phone. In addition, IBM [1] has successfully developed an enterprise speech solution, named IBM WebSphere Voice Server, which provides voice-enabled applications to give the customers, employees and suppliers more flexible access to information and services. In Vietnam, current services provided by the contact centers are mostly under-run by manpower or through an SMS protocol. In 2010, R&D group from AILab has proposed a Vietnamese spoken dialog system for the inquiry of stock information over the phone with the best accurate rate of 87.3% [3]. However, the system was just built to process only stock ticker symbols and users were not required to speak naturally. Since then there was no application of voice server in the Vietnamese industries, and its research is still on hung.

Meanwhile, in developing countries, nearly 1.5 billion people live without electricity [13] and 752 million are illiterate [14] – two constraints that make accessing information challenging. To exacerbate this problem, the majority of these people live in rural areas [13], which are often hard to reach because of inadequate roads. Information about farming techniques is particularly important because agriculture is a major source of livelihood for most rural people though they often rely on rudimentary methods [15]. To enhance the agricultural life, one must ensure the flow of information to farmers; that is to bridge an agricultural info channel to farmers effectively. Putting on this mission, we have the statement of agricultural extension [16].

Conventional approaches for agricultural extension, like extension workers and infomediaries, serve their roles well. But the current trend of ICT services is rising and proven to be more efficient [15]. As SLP technologies advance, the IVR systems are greatly enhanced, allowing for natural speaking style, domain adaptability, and robust speech recognition [9]. Thus, taking the ripe fruit, we brought our IVR framework to provide an automated information channel for agricultural extension – that is, an automatic call-center answering questions on farming techniques. Figure 1 illustrates a typical dialog session between a farmer and the system. With the agriculture domain, its ASR engine needs to be rebuilt in order to maintain a sustainable recognition performance. This involves in collecting corpora for the applied domain and adjusting the models. However, after all the hard tasks, recognition accuracy only satisfies a feasible level; errors remain quite high. The problem originates from the amount of training/adapting data available.

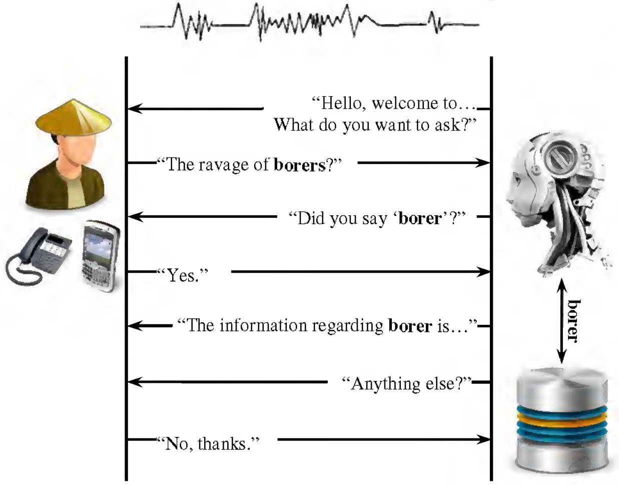

"Hello, welcome to…
What do you want to ask?"
"The ravage of **borers**?"
"Did you say '**borer**'?"
"Yes."
"The information regarding **borer** is…"
"Anything else?"
"No, thanks."
borer

Figure 1. An agricultural session of human-machine dialog

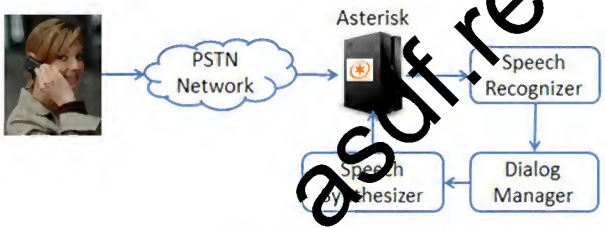


Figure 2. System design

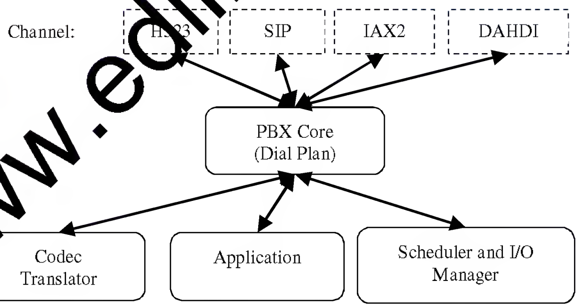


Figure 3. Overview of Asterisk architecture

Insufficient data would definitely lead to the degradation of ASR performance. In popular languages like English and Mandarin, there exist a large amount of speech corpus resources for system development. But for many dialectical variants and languages such as Malay or Vietnamese, this is not the case. Unfortunately, the required effort to develop speech corpora is both costly and time consuming. Furthermore, large scale acquisition campaigns are simply not feasible. These languages are referred to by the term "under-resourced languages" [7]. Amongst state-of-the-art techniques, Subspace Gaussian Mixture Model [11] has proved to be effective against under-resourced circumstances. Thus we resort to its mechanism for powering the IVR system's ASR component, targeting on agricultural extension service. This paper focus on refining the work of [9], altering its ASR and TTS engines to comply with the under-resourced condition. Section II presents the system architecture and its components, while Section III gives experimental results. Finally, Section IV concludes the paper.

II. The IVR System and its Components

This section describes the Agricultural IVR system. It is responsible for answering incoming calls of a specific agricultural query. Figure 2 illustrates the four main modules composing the system: an Asterisk PBX Server, a speech recognizer, a speech synthesizer, and a dialog manager. The Asterisk server manages telephone signal transmissions between users and the system over PSTN network, while dialog manager executes the tasks of query and processing information. Both the speech recognizer and synthesizer operate as a communication layer by dealing with speech-to-text conversions and vice versa. Incoming queries will then be interpreted and responded appropriately.

Each of the following subsection will describe the system's components and their underlying technologies.

A. Asterisk

Asterisk is a free software implementation for telephone private branch exchange that transforms a computer into a communication server and can be used as a telephony engine and application toolkit. It is also a framework that allows selection and removal of particular modules, allowing us to create a custom telephony system [4]. Asterisk's well-thought-out architecture gives flexibility for creating custom modules that extend our phone system, or even serve as drop-in replacements for the default modules. Asterisk flexibility allows it to be deployed as PBX, VoIP, IVR as well as Voice Mail system [5]. A PBX is a system which allows one telephone to make connection with other telephone and telephone services. Generally, it can interoperate with almost all-standards-based telephony equipment using comparatively inexpensive hardware which makes it easier to connect with traditional telephony network as well as various computer networks. This way, Asterisk PBX server can be added with a couple of new functionalities. The new functionalities can be added by writing dial plan scripts in some Asterisk's own extension languages or by including custom loadable modules written in C or by implementing the Asterisk Gateway Interface (AGI) programs using any programming language like Perl, python, shell scripts, etc. [6]. Figure 3 depicts the Asterisk architecture.

The heart of any Asterisk system is the PBX core. It is the essential component that takes care of bridging calls. The core also takes care of other items like codec translator, scheduler and I/O manager, application, and other modules.

B. Speech Recognizer

To cope with the problem of limited training data, Subspace Gaussian Mixture Model (SGMM) acoustic modeling techniques [11] are used. In contrast to the usual approaches that deploy a set of universal phones to cover multiple languages, the approach of SGMM uses distinct phone sets but shares a large amount of parameters across languages. In SGMM, HMM-states' feature distributions are Gaussian Mixture Models (GMMs) with a common structure, constrained to lie in a subspace of the total parameter space. The parameters that define this subspace can be shared across languages/domains. Formally defined, the feature distribution of a HMM-state j is given by:

$$p(x \mid j) = \sum_{i=1}^{I} w_{ji} \mathrm{N}(x; \mu_{ji}, \Sigma_i) \qquad (1)$$

where x is the feature vector and N(x; μ, ∑) is the Gaussian function. This might look a little similar to the conventional GMM, however, the difference lies in the way of representing mixtures. An intuitive illustration for both models can be seen in Figure 4. For SGMM, a particular state j is associated with a vector v_j which determines the means and weights as follows:

$$\mu_{ji} = \mathrm{M}_i v_j \qquad (2)$$

$$w_{ji} = \frac{\exp w_i^T v_j}{\sum_{i'=1}^{I} \exp w_{i'}^T v_j} \qquad (3)$$

where M_i and w_i are shared across all state distributions. In addition, the covariance matrices $\sum_i$ are globally shared as well. Together, M_i, w_i and $\sum_i$ form the set of globally shared parameters, as opposed to the state-specific vectors v_j.

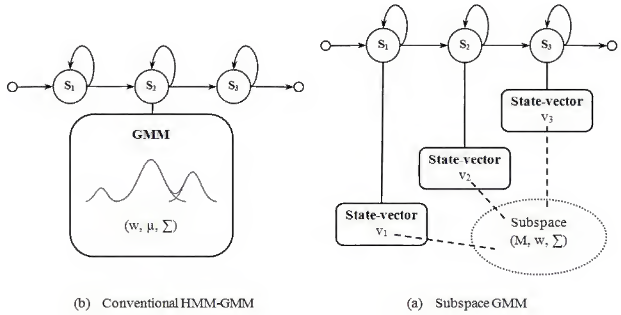
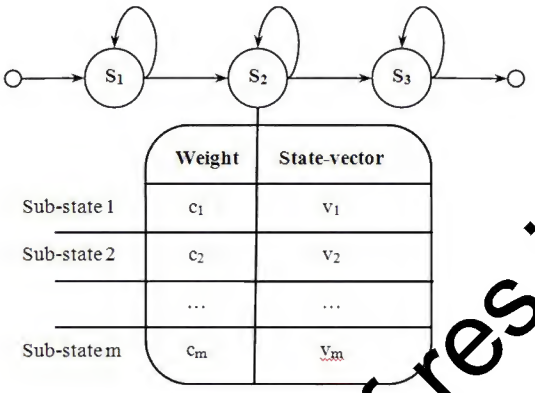


Figure 4. HMM structures.



	Weight	State-vector
Sub-state 1	c₁	v₁
Sub-state 2	c₂	v₂
	…	…
Sub-state m	cₘ	vₘ

Figure 5. SGMM with sub-states.

To achieve a balance between the amount of shared and state-specific parameters, the notion of a "sub-state" [11] was introduced. Instead of just one state-vector, the feature distribution for a state can be represented by a mixture of M vectors, each with its own weight c. Figure 5 gives a clearer picture on this notion. In this case, the feature distribution of a state j is given by:

$$p(x\,|\,j)=\sum_{m=1}^{M_j} c_{jm}\sum_{i=1}^{I} w_{jmi}\mathrm{N}(x;\mu_{jmi},\Sigma_i) \qquad (4)$$

$$\mu_{jmi}=\mathrm{M}_i v_{jm} \qquad (5) \qquad\qquad w_{jmi}=\frac{\exp w_i^T v_{jm}}{\sum_{i=1}^{I}\exp w_{i'}^T v_{jm}}. \qquad (6)$$

Utilizing SGMM, one can deal with the problem of limited training data for under-resourced condition. Indeed, the set of globally shared parameters {M$_i$, w$_i$, Σ$_i$} can be trained on out-of-domain data, while the state-specific vectors {v$_j$} can be trained on a limited amount of in-domain data. In the experiments for this paper, broadcast news and agricultural telephony are selected as the targets for well-resourced and under-resourced domain respectively (i.e., the broadcast news corpus serves as the out-of-domain data and the agricultural telephony corpus serves as the in-domain data).

C. Speech Synthesizer

The original work [9] employs VOS's corpus-based version [8] to power its TTS engine. This has the advantage of naturalness and intelligibility, but suffers from the oversize database (i.e., more than 4 GB/40h duration) and therefore lack of portability and dialect variations. In cases of under-resourced conditions, even several hours of speech are unaffordable, let alone 40h/4GB. Building a corpus-based TTS engine would therefore infeasible.

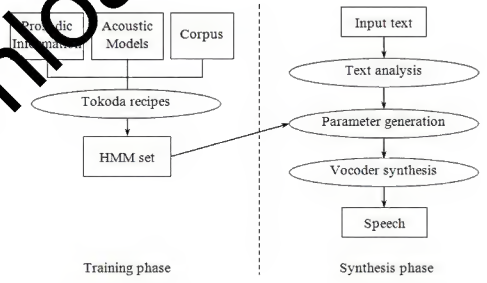


Figure 6. HMM-based speech synthesis.

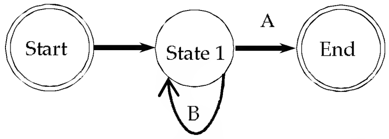


Figure 7. Keyword spotting FSM.

Complying with the condition, we adapt the TTS core to parametric synthesis – the HMM-based synthesis [12] which has the advantages of lightweight storage and smooth prosody. Figure 6 gives an outline view on the TTS flow. Required data can be as low as 45 minutes of speech and highly natural voice could be achieved at ~2 hour level. Sample speech for training the models was collected from three candidate speakers of Saigon, Hue, and Hanoi dialects, providing 3 different voices for user options – the first Vietnamese TTS system to be capable of.

D. Dialog Manager

Acting as the brain of an IVR system, dialog manager controls the calling sessions. A preset communication script is enforced on each session. It starts with a welcome message and waits for user response. Users will ask an agricultural question in a very natural sense. The dialog manager is bound to find out and speak back an appropriate answer. Users could also choose to ask another question from this step until an end is confirmed.

In contrast to voice commands by keywords, our system provides a flexible means for voice communication by natural language interaction. One can ask a question as they would do to a person. By saying something such as "What is borer?" or "How can I fight locust?" or simply "The ravage of borers?" the dialog manager will fetch back the appropriate answer. To achieve this goal, a keyword spotting mechanism is proposed to pick out important terms from a complete sentence. Let A be the set of agricultural keywords and B stand for the set of grammar terms. The finite state machine (FSM) depicted in Figure 7 is used to render the keyword spotting functionality. In this sequence, agricultural terms (A) are always required for queries while grammar terms (B) are optional and can be disposed of. If the "end" state is not reached, a null query will be assumed.

III. Experiments

This Section focuses on the evaluations of the ASR engine, IVR online trial, and runtime response. All of them are conducted on the dataset described below.

A. Datasets

TABLE I. Speech Corpora

Corpus	***Duration***	***#Speakers***
VOH	27 hours	18
Agricultural Telephony (AT)	7.2 hours	62

We first collect the agricultural telephony speech corpus from 62 speakers of Mekong Delta which represent for the farmer dialects. Total duration is roughly 7.2 hours with a vocabulary size of 103 words (including keywords, grammar terms and confirmation words) several of which are listed in Table II. Next, we compile it together with the VOH corpus to evaluate the recognizer. Both corpora are converted to an identical format of 16 KHz, 16 bits, mono. They are further parameterized into 12-dimensional MFCC, energy, plus their delta and acceleration (39 length front-end parameters).

TABLE II. Lexicon Samples

bạc	màu	châu	chấu	cuốn
lá	cháy	vàng	úng	giống
nhiễm	khuẩn	làm	đòng	ngập
trắng	sâu	đốm	vòng	thân
cho	ừ	ok	tôi	hãy
sai	đúng	rồi	vâng	không

The corpora (shown in Table I) are then divided into subsets for training and testing 2 target ASR engines (i.e., the baseline and SGMM systems). Table III summarizes the training and test sets devised for experiments.

TABLE III. Training and Test sets

	Training		**Test**	
	hours	***corpora***	***hours***	***Corpora***
Baseline	6.2	6.2h Agriculture (AT)	1	1h AT
SGMM	33.2	27h VOH + 6.2h AT	1	1h AT

Language models (trigrams) for the recognizers are built by interpolating individual models trained from the Web text corpus and the training data's transcriptions.

B. Transcription Evaluation

TABLE IV. Transcription performances

	Baseline	**SGMM**
%WAR	90.26%	93.04%

In this experiment, the recognizers are evaluated on the task of speech transcription. Performances are reported for two different systems: baseline and SGMM. The baseline recognizers are based on conventional 3-state left-to-right HMM triphone models, with 18 Gaussians per state. The SGMM system's shared parameters are estimated using data from both VOH and AT, while the state-specific parameters are trained on AT data only. An SGMM configuration with 400 shared Gaussian components (I = 400), 40-dimensional state-vectors and 12 sub-states per state is used.

Table IV summarizes the performances of the recognizers. Using SGMM, an absolute improvement of 2.78% WAR over the baseline is achieved. The results confirm the benefit of SGMM in taking advantage of resources in other domains whenever only a small amount of training data is available.

C. IVR Online Trial

For online trials, the whole IVR system was deployed in a real data center which connected to a telephone network. Users use their mobile phone to dial the system number and interact with our voice server. We ask 30 volunteers each to make 10 calling attempts separately. That means users don't need to get used to the system and they can feel to speak whatever they want in terms of agricultural query. Each calling session is processed by both ASR engines (i.e., baseline and SGMM) simultaneously. Results, in accuracy rates, can be seen in Figure 8.

As expected, SGMM gains the upper hand over the baseline, but subdued to its own transcription performance. For losing utterance constraints in online trials, the score decreased approximately 2.62% when compared to transcription tests. However, the average WAR of 90.42% in the online trial indicates that our proposed system could be an effective call center for agricultural extension services.

D. Runtime response

As an agricultural extension service, its response time is crucial. In order to be deployed, response timing must be real-time equivalent or even better. This experiment measures the running time for each communication session, including both ASR and TTS computations. The same 30 volunteers who participated in the online trials are asked to communicate with the server using random utterances. Processing durations are logged and an average response time of 2.349 seconds can be derived. Figure 9 plots the timing performances of the first 50 loops.

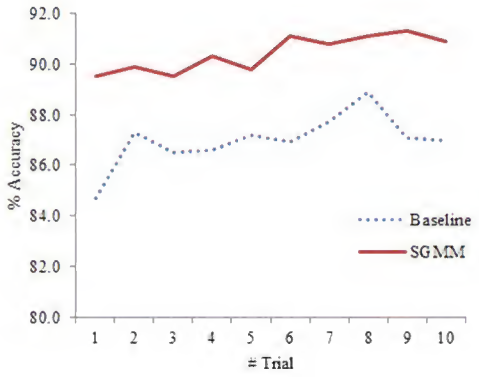


Figure 8. IVR performances.

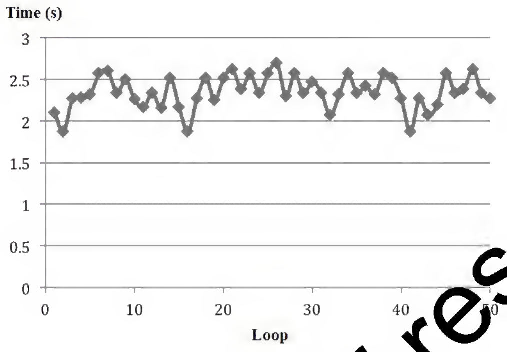


Figure 9. Runtime performances.

IV. Conclusion

This paper has presented a critical enhancement for voice server on under-resourced conditions. Should we run low on corpora, SGMM-based ASR and HMM-based TTS techniques are always there for us. Experimental results did confirm the hypothesis.

Acknowledgment

This work is part of the "iSago" project, supported by the HCM Department of Science and Technology.

References

1. Steve Cawn, Baiju Mandalia, Wendi Nusbickel, Incorporating IBM speech solutions into a service oriented architecture, IBM WebShpere Voice Server WhitePaper, June, 2006 .
2. V. Zue et al, "JUPITER: A Telephone-Based Conversational Interfacefor Weather Information", IEEE Transactions on Speech and Audio Processing, Vol 8, No. 1, Jan 2000, pp.85–96.
3. Quan Vu et.al "A Spoken Dialog System For Stock Information Inquiry," IT@EDU, 2010.
4. Jim Van Meggelen, Leif Madsen, and Jared Smith, "Asterisk™: The Future of Telephony, Second Edition",2007.
5. David Gomillion, Barrie Dempster, "Building telephony systems with asterisk", Packet Publishing, 2005.
6. Md. Zaidul Alam, Saugata Bose, Md. Mhafuzur Rahman, Mohammad Abdullah Al-Mumin, "Small office PBX using Voice over IP" in Proc. IEEE ICACT 2007, Feb 12-14 2007.
7. Q. Vu et al., "Progress in transcription of Vietnamese broadcast news," International Conference on Communications and Electronics (ICCE'06), pp. 300-304, Oct. 2006.
8. Q. Vu et al., "VOS: the corpus-based Vietnamese text-to-speech system," Research, Development and Application on Information and Communication Technology, 2010.
9. Q. Vu et al., "A Robust Vietnamese Voice Server for Automated Directory Assistance Application, " VLSP 2012.
10. Le, V. B., and Besacier, L., "Automatic speech recognition for under-resourced languages: application to Vietnamese language," IEEE Transactions on Audio, Speech & Language Processing, vol. 17, iss. 8, pp. 1471–1482, 2009.
11. Povey, D., et al., "Subspace Gaussian mixture models for speech recognition," Proceedings of ICASSP'10, 2010.
12. K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," IEEE Speech Synthesis Workshop, 2002
13. G. Legros, I. Havet, N. Bruce, and S. Bonjour, "Energy access situation in LDCs and Sub-Saharan Africa" in The Energy Access Situation in Developing Countries: A Review focusing on the Least Developed Countries and Sub-Saharan Africa, New York: UNDP & WHO, 2009, ch. 3.
14. United Nations Education, Scientific and Cultural Organization "Education for All Global Monitoring Report 2010, Reaching the marginalized" UNESCO, Paris, France, 2009.
15. Clifford Schmidt et.al, "Impact of Low-Cost, On-Demand, Information Access in a Remote Ghanaian Village," ICTD2010, December 13–15, 2010, London, U.K..
16. A. Maunder, Agricultural Extension: A Reference Manual (1st Edition), FAO, 1973.